FACT: FIDELITY-AWARE CONSTRUCTION OF ARTICULATED TWINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visually plausible articulated assets may still fail during contact interactions or exhibit inaccurate motion. We present **FACT** (Fidelity-Aware Construction of Articulated Twins), an agentic framework that progressively constructs articulated twins to improve geometry, contact, and dynamic fidelity. The agent drives an evidence–diagnosis–revision loop on a shared editable representation, selecting measurements and model edits using quantitative feedback, while numerical tools execute and validate the updates. It reconstructs editable articulated geometry from images through feature planning, targeted measurements, and diagnostic refinement. On this reference, it repairs collision proxies through task-aware local repartitioning before fidelity-constrained compression. Finally, it constructs response models from passive-response videos, using simulation residuals to guide model revision and constrained physical parameter fitting. Experiments show that FACT improves geometric reconstruction over baselines, enables more reliable interaction with simpler collision proxies, and better reproduces held-out physical responses than direct parameter inference.

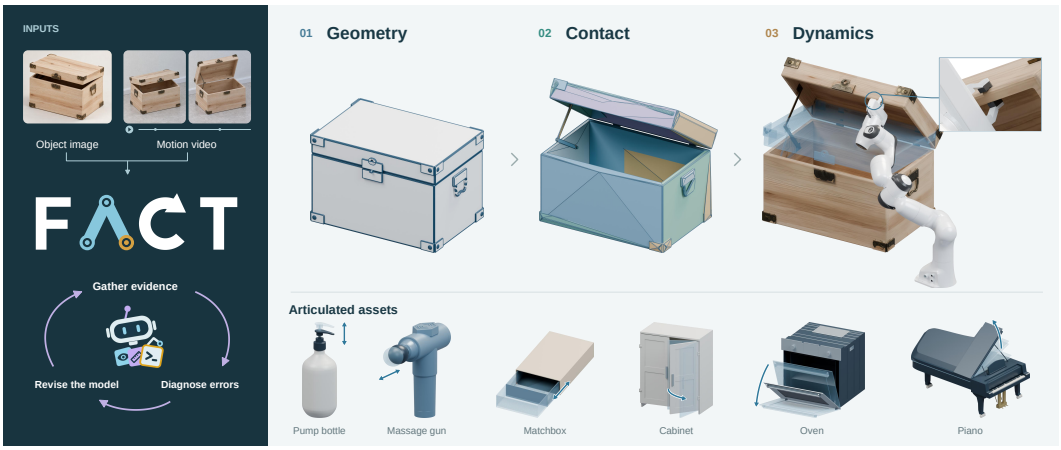

Figure 1: **Geometry, contact, and dynamics. FACT** reconstructs editable articulated objects, refines collisions to preserve free space for interaction, and calibrates motion dynamics from video.

1 INTRODUCTION

Simulation enables embodied agents to gain experience through repeated interactions with diverse objects. Advances in simulation platforms, articulated asset collections, and procedural generation have broadened this opportunity (Xiang et al., 2020; Geng et al., 2023; Joshi et al., 2025), while generative and agentic systems make asset creation from visual inputs increasingly accessible (Le et al., 2025; Zhou et al., 2026; Wang et al., 2026b). Yet increasing the number of visually plausible articulated assets does not by itself provide more faithful interaction experience. An asset determines both what a robot observes and what happens when it acts: where contact can be established, which motions are feasible, and how the object's motion evolves over time. Building useful simulation-ready assets therefore requires capturing interaction feasibility and motion responses alongside shape and articulation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This gap is particularly clear in automatic articulated asset creation. Articraft generates and refines programs for parts and joints from images (Zhou et al., 2026). However, image-based reconstruction can still misrepresent drawer depth, handle clearance or interior cavities, yielding an asset that looks plausible but does not support the intended interaction. Even when these structures are accurate, collision approximation can compromise otherwise feasible interactions. Pipelines such as EmbodiedGen V2 use CoACD to obtain convex collision proxies (Wang et al., 2026b; Wei et al., 2022). Although CoACD accounts for collision-related geometric error, proxies can still fill handle openings or obstruct drawer travel. Yet contact feasibility alone does not ensure a faithful motion response: Articraft uses language-model priors for mass and damping, whereas reproducing instance-specific responses calls for calibration against observed trajectories. These considerations motivate three complementary aspects of fidelity: *geometry fidelity* in structure and dimensions, *contact fidelity* in task-specific interaction and motion feasibility, and *dynamic fidelity* in motion responses.

Addressing these errors calls for progressively grounding a shared articulated model. Reconstructed geometry and articulation guide collision refinement and, together with the refined proxies, form a fixed basis for dynamic calibration. The appropriate measurement and correction depend on the part and the failure: a narrow opening requires different evidence from an incorrect joint response. An agent drives this evidence–diagnosis–revision loop, selecting measurements and model edits using semantic understanding and quantitative feedback, while numerical tools execute and validate the updates. The central design is to link geometric discrepancies to editable features, free-space and motion violations to collision partitions, and trajectory residuals to response models and physical parameters.

We present **FACT**, an agentic framework that progressively constructs and calibrates an articulated digital twin (Fig. 1). First, the agent plans parametric features and selects measurements to constrain their dimensions and placement, using local diagnostics to reconstruct editable geometry and articulation from an image-generated mesh. Keeping this geometry and articulation fixed, it identifies likely interactions and revises local collision partitions to repair free-space and motion violations, then compresses the proxies under fidelity constraints. Finally, with the resulting geometry, joints, and proxies fixed, the agent uses videos to estimate trajectories in the recovered joint coordinates and construct response models. It then uses simulation residuals to revise the models and guide constrained parameter fitting.

Our contributions are threefold:

- We introduce an agentic reconstruction method that combines feature planning, targeted measurements, and diagnostic refinement to recover editable articulated geometry.
- We develop agent-guided collision refinement that diagnoses task-specific contact failures, revises local partitions, and compresses proxies under fidelity constraints.
- We formulate residual-guided dynamic calibration in which the agent autonomously constructs and revises response models and directs constrained system identification from interaction observations.

2 Related Work

Articulated object reconstruction and generation. Articulated reconstruction recovers part geometry and kinematics from observations across object states or continuous motion (Jiang et al., 2022; Liu et al., 2023; Weng et al., 2024; Liu et al., 2025b; Ai et al., 2026), while geometry-based articulation infers part structure and joints from meshes or point clouds (Qiu et al., 2025; Wang et al., 2026a; Li et al., 2026). Generative approaches learn joint representations of geometry, connectivity, and motion (Lei et al., 2023; Liu et al., 2024; Gao et al., 2025; Su et al., 2025), with image-conditioned methods producing articulated assets from visual cues (Chen et al., 2024; Liu et al., 2025a). Code-based and agentic pipelines express articulated assets as executable models (Zhao et al., 2025; Le et al., 2025; Zhou et al., 2026). Complementary work addresses hidden geometry completion (Boudjoghra et al., 2026), non-penetration and mobility constraints (Kreber & Stueckler, 2025), and physical attributes for simulation (Cao et al., 2026; Yang et al., 2026b).

Programmatic mesh reconstruction and agentic 3D modeling. Structured geometric representations range from compact assemblies of primitives and superquadrics (Ye et al., 2025; Fedele et al.,

2025; Tavernini et al., 2026) to shape programs built from parameterized parts, assembly relations, and reusable operations (Jones et al., 2020; 2021; 2023; 2026). Shape-to-code approaches reconstruct executable programs from point clouds or images (Dai et al., 2026; Rukhovich et al., 2025; Kolodiazhnyi et al., 2025). Program generation is complemented by agentic frameworks that combine structural planning, code execution, and validation (Zhang et al., 2026; Zhou et al., 2026; Lin et al., 2026), and by visual or geometric feedback for iterative refinement (Kabisov et al., 2026; Hu et al., 2026). In **FACT**, feature plans guide measurements of dimensions and placement, while local diagnostics determine whether to revise a feature representation or its parameters.

Collision geometry for interactive simulation. Collision proxies can be constructed through approximate convex decomposition of non-convex meshes (Mamou et al., 2016; Wei et al., 2022). Decomposition strategies include learned cutting policies, visibility-based partitioning, and neural feature fields (Luo et al., 2025; Fokin & Savva, 2026; Yang et al., 2026a). Application requirements also guide decomposition: methods preserve navigable space (Andrews, 2024) or maintain coherent decompositions across animation poses (Thul et al., 2018). **FACT** uses CoACD as a backend, with task and motion diagnostics guiding local repartitioning and subsequent compression under fidelity constraints.

Physical digital twins and system identification. Visual system identification couples differentiable simulation and rendering to infer physical parameters from videos (Jatavallabhula et al., 2021; Li et al., 2023), while deformable digital twins recover geometry and material response from observed interactions (Zhong et al., 2024; Jiang et al., 2025). Robot-interaction pipelines instead use torque sensing, proprioception, and motion observations to estimate object properties (Pfaff et al., 2025; Chen et al., 2025; Lou et al., 2026). RigPI combines visual priors with torques and poses to identify inertial and frictional parameters of articulated objects (He et al., 2026). In **FACT**, the agent autonomously constructs and revises response models from observed motion and directs constrained numerical fitting.

3 METHOD

3.1 OVERVIEW

FACT constructs an articulated twin from object images and motion videos (Fig. 2). We represent the twin as $\mathcal{A} = (G, K, C, \theta)$: part geometry, a kinematic graph, collision proxies, and physical parameters. Feature planning, measurement, refinement, and articulation establish (G, K) (Sec. 3.2). Task-aware contact refinement repairs C on fixed (G, K) before compression under volumetric, free-space, and sampled-motion constraints (Sec. 3.3). Dynamic calibration constructs and revises a passive response model and fits θ to video trajectories on fixed (G, K, C) (Sec. 3.4). Each stage follows an evidence–diagnosis–revision loop: the agent selects measurements and model revisions, while numerical tools execute these operations and return quantitative feedback. Shared part identities and body frames maintain correspondence across geometry, collision proxies, and dynamic.

3.2 GEOMETRY FIDELITY: FEATURE PLANNING, MEASUREMENT, AND REFINEMENT

We separate feature planning from numerical measurement, using reconstruction feedback to revise the feature representation or its parameters.

Feature planning. The agent inspects whole-object and isolated part views of an image-generated mesh M to identify logical parts and their structural requirements. A feature plan specifies feature types, relations, parameters to measure, and links to agent-identified source regions. Each part may combine extrusions, sweeps, shells, and subtractive features, with custom constructions for shapes outside this vocabulary. Features are associated with the visible structures they must preserve, including openings, contacts, and repeated elements. Planning determines what to represent and measure, leaving numerical values and measurement procedures to the next step.

Feature-conditioned measurement. For each planned parameter, the agent selects a source region, coordinate frame, and measurement procedure; geometric tools compute the values. Projected contours and axial extents constrain an extrusion's profile and depth, whereas cross-section centers

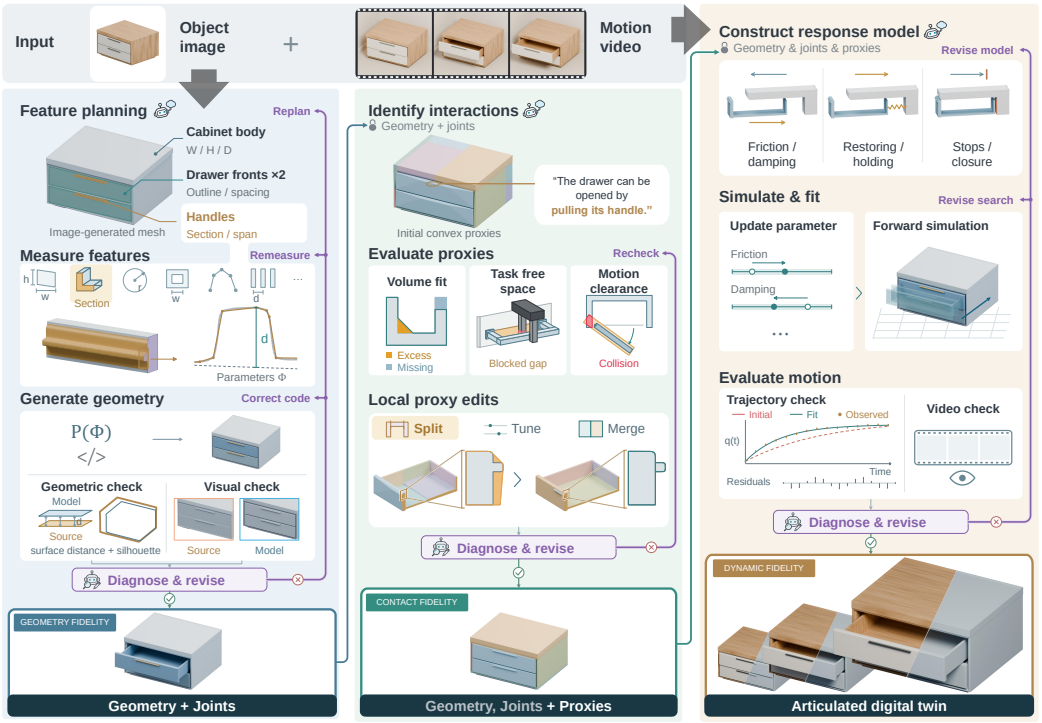

Figure 2: **FACT overview.** The agent drives an evidence–diagnosis–revision loop on a shared articulated representation. It plans geometric features and selects measurements, directs task-aware collision repair followed by fidelity-constrained compression, and constructs and revises response models while guiding constrained parameter fitting.

and contact regions constrain a handle's path and endpoints. Measured contours are reduced to editable profile controls. Dependent parameters are derived from measurements using symmetry, repetition, and contact relations. We retain each parameter's measurement or derivation for later revision. The agent implements the plan as an executable program $P(\phi)$ with geometric parameters ϕ, which generates reconstructed part meshes $\widehat{G}$ from editable features rather than copied source triangles.

Diagnostic refinement. We compare $\widehat{G}$ with M using bidirectional surface distances, matched-view silhouettes, and side-by-side renders. Part-level diagnostics use agent-identified source correspondences to localize errors, while targeted views or sections help resolve structures obscured in standard views. The agent traces discrepancies to the responsible modeling decision: an inadequate feature representation triggers replanning, incorrect dimensions or placement trigger remeasurement, and implementation errors trigger program correction. The revised program regenerates the meshes for reevaluation. A reconstruction is accepted only when it passes both automated geometric validation and the agent's visual inspection.

Functional completion and articulation. After validation, a separate pass adds or refines minimal functional structures, such as cavities, supports, and mounts. Category knowledge guides unobserved structures, while measured envelopes and functional clearances constrain their dimensions. Parts that move together are grouped into rigid bodies, and each joint in K specifies its parent and child bodies, type, axis, origin, and motion limits. Sampled poses, including travel endpoints, reveal interferences that prompt geometric or joint-placement corrections. The resulting G and K form the articulated reference for contact refinement.

3.3 Contact Fidelity: Task-Aware Collision Refinement

We repair task-dependent collision errors before compressing the proxy under fixed quality constraints.

Collision representation and initialization. For each rigid body b, let G_b denote the region occupied by its reference geometry. Its collision proxy is $C_b = \bigcup_{h=1}^{H_b} C_{b,h}$, a union of H_b convex hulls. We use C to denote the full collection of hulls across bodies. Decomposition regions serve as inputs to CoACD (Wei et al., 2022) and may be repartitioned within a body while retaining all source geometry. Reference geometry G and articulation K remain fixed. A complete default regional decomposition C^0 initializes the proxy and sets the total hull budget $H_{\max} = H(C^0)$.

Task-aware diagnostics. The agent identifies likely interactions and selects contact regions, access spaces, and motion clearances for evaluation. Task free space is defined outside the union of all reference body geometries at each pose, independently of the candidate proxy. We evaluate per-body volumetric overlap, missing and excess volume, task-region occupancy, and false interbody collisions at reference-confirmed collision-free poses. Motion checks sample the full range, including endpoints, with denser sampling near failures; diagnostics identify the responsible hull pairs and source regions.

Diagnostic partition refinement. Based on these diagnostics, the agent selects a local edit: repartitioning or splitting a region, adjusting regional decomposition parameters, or merging compatible regions or hulls within a body. Separating a handle into side supports and a bridge, for example, can prevent convex hulls from filling its opening. Geometric tools execute the edit and regenerate affected proxies, followed by whole-object reevaluation. Refinement seeks a proxy within the hull budget that passes the sampled motion checks while bounding increases in volumetric and free-space errors relative to C^0. Once found, this proxy is fixed as the quality reference C^a.

Feasibility-first compression. Starting from C^a, the agent searches through local edits for a simpler proxy, prioritizing total hull count H over total vertex count V:

$$\begin{aligned} \operatorname*{lexmin}_{C} \quad & \big(H(C), V(C)\big) \\ \text{subject to} \quad & C \in \mathcal{F}(G,K), \quad H(C) \le H_{\max}, \\ & \mathbf{e}(C) \preceq \mathbf{e}(C^a) + \boldsymbol{\delta}. \end{aligned} \tag{1}$$

Here, $\mathcal{F}(G,K)$ contains valid proxies with no self-collisions over the articulated motion space. The error vector $\mathbf{e}$ collects per-body $1-\mathrm{IoU}$, missing and excess volume ratios, and free-space occupancy in each selected region. The tolerance vector $\boldsymbol{\delta}$ bounds the componentwise error increases relative to C^a. An edit is accepted only if it improves the lexicographic objective and satisfies these constraints.

3.4 Dynamic Fidelity: Video-Based Calibration

Given one or more videos, we calibrate an object's dynamics with the reconstructed articulated twin (G, K, C) fixed.

Motion-grounded parameterization. The recovered kinematic model defines the joint types and coordinate system used to represent the observed motion. Within this representation, the agent selects motion-estimation procedures and invokes analysis tools to estimate time-aligned joint trajectories and initial states. It uses the recovered mechanism, visual cues, and observed motion to autonomously construct the response model, specifying its functional form and initializing its physical parameters θ. Depending on the mechanism, θ may include parameters governing damping, friction, and state-dependent restoring or holding terms. The parameterization preserves the physical relationships among these quantities and the reconstructed geometry.

Simulation-based inference. For observed trajectories $\{q_k^{\mathrm{obs}}\}_{k=1}^m$, $m \ge 1$, forward simulation predicts $q_k^{\mathrm{sim}}(\theta; \xi_k)$, where ξ_k denotes sequence-specific initial conditions. The agent defines a constrained search space Ω, selecting which physical parameters and uncertain initial-state components to optimize, together with their admissible bounds. Within each fitting step, the response-model form remains fixed, and unselected quantities retain their current estimates.

$$\min_{(\theta, \{\xi_k\}) \in \Omega} \sum_{k=1}^{m} \left[\mathcal{L}_q\big(q_k^{\mathrm{sim}}(\theta; \xi_k), q_k^{\mathrm{obs}}\big) + \mathcal{R}_k(\xi_k) \right]. \tag{2}$$

Here, $\mathcal{L}_q$ measures time-aligned trajectory discrepancies, while $\mathcal{R}_k$ constrains any optimized initial-state components around their observed estimates; it is omitted when initial conditions are fixed.

Physical parameters are shared across sequences of the same joint and configuration. The numerical solver returns fitted values, simulated trajectories, and residuals.

Residual-guided refinement. The agent uses residual patterns to revise the response-model form and the optimization variables and bounds in Ω. Speed-dependent discrepancies prompt examination of damping, while incorrect stopping or holding behavior prompts examination of resistance terms. When discrepancies suggest uncertain observations or initialization, the agent rechecks the source video or revises which initial-state components are fitted. Numerical tools solve the revised problem, and forward simulation evaluates the resulting motion. After calibration, the response model and physical parameters are fixed for evaluation on held-out clips.

Table 1: **Quantitative comparison of geometry fidelity.** Gray rows show upstream Rodin meshes for reference; bold marks the best reconstruction results within each setting.

Method	CD↓	SD-P95↓	F@1%↑	F@2%↑	Avg. SIoU↑	Min. SIoU↑
Image inputs						
Codex (GPT-6)	1.1238	3.6060	0.6148	0.8237	0.8813	0.8310
Articraft (GPT-6)	1.3117	3.3719	0.5687	0.7615	0.8597	0.8068
FACT (GPT-6; original)	1.0821	**2.7884**	0.6838	0.8391	**0.9381**	**0.9166**
FACT (GPT-6; parts)	**1.0317**	2.9972	**0.6937**	**0.8519**	0.9317	0.9051
Rodin (original)	1.6840	1.9756	0.5663	0.7206	0.9393	0.9164
Rodin (parts)	1.4353	2.6555	0.6041	0.7703	0.9336	0.9075
Oracle mesh inputs						
CAD-Recode	1.2073	3.1952	0.6164	0.7990	0.8596	0.7475
MeshCoder	3.4268	6.9213	0.3541	0.5018	0.6366	0.5035
PrimitiveAnything	1.4028	4.7916	0.5354	0.7520	0.8241	0.7674
SuperFlex	1.9266	3.2654	0.3937	0.6578	0.8852	0.8455
Codex (GPT-6)	0.5402	1.0572	0.8783	0.9291	0.9888	0.9790
FACT (GPT-6)	**0.0874**	**0.3740**	**0.9899**	**0.9990**	**0.9902**	**0.9861**

4 EXPERIMENTS

4.1 GEOMETRY FIDELITY

Experimental setup. We evaluate geometry reconstruction on 21 objects under image-input and oracle-mesh settings. In the image-input setting, we compare **FACT** against Codex and Articraft (OpenAI, 2025; Zhou et al., 2026). **FACT** reconstructs from either original or part meshes generated by Rodin Gen-2.5 (Zhang et al., 2024; 2025). For oracle inputs, reference meshes replace generated geometry to isolate reconstruction quality from upstream 3D generation. We compare against CAD-Recode, MeshCoder, PrimitiveAnything, SuperFlex, and Codex (Rukhovich et al., 2025; Dai et al., 2026; Ye et al., 2025; Tavernini et al., 2026). Learned baselines operate on whole-object geometry, whereas agentic methods can inspect the provided mesh structure. All agent-based methods use GPT-6 Astra (OpenAI, 2026b) with High reasoning level, with Codex using its default general-purpose workflow.

Evaluation protocol. Outputs are evaluated against the complete reference geometry in an aligned canonical pose. We report symmetric Chamfer distance (CD), the 95th-percentile reconstruction-to-reference surface distance (SD-P95), surface F-scores, and silhouette IoU (SIoU). CD and SD-P95 are normalized by the reference bounding-box diagonal and reported as percentages. Scores are averaged equally across objects.

Reconstruction accuracy. With oracle inputs, **FACT** leads all six metrics and reduces CD by 83.8% relative to Codex (Tab. 1), supporting the value of structured reconstruction beyond access to accurate geometry alone. Both image-based variants also outperform Codex and Articraft, while improving CD and surface F-scores over their upstream meshes. Fig. 3 complements these aggregate results with matched-view comparisons and enlarged local structures. Fig. 3 further shows that **FACT**'s gains extend beyond overall shape agreement to more accurate local structures and dimensions.

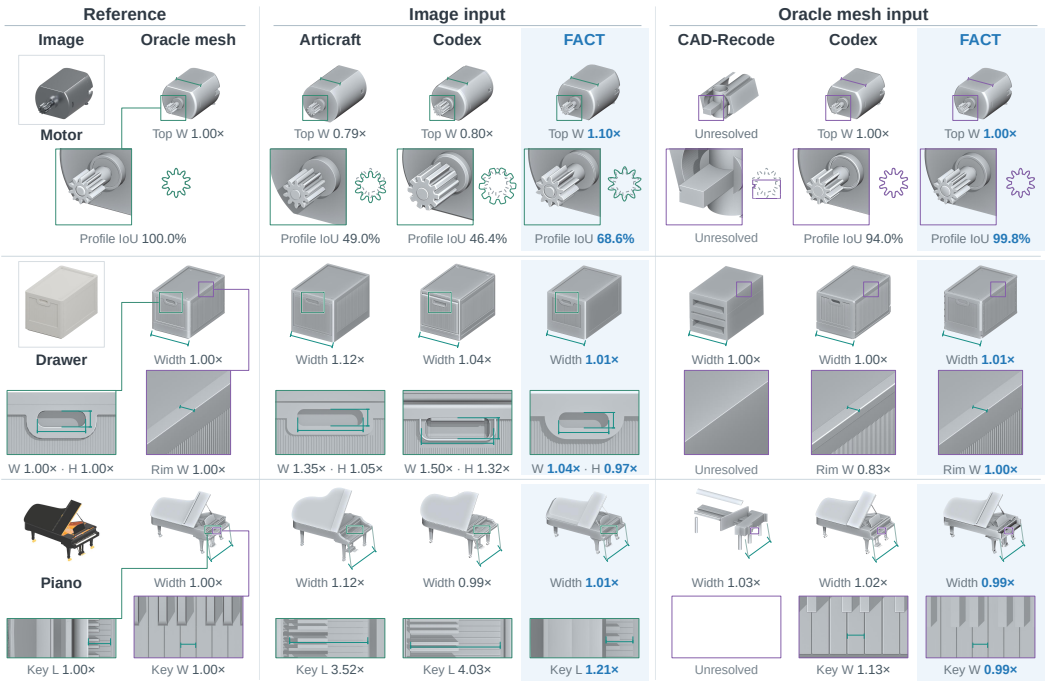

Figure 3: **Geometry fidelity. FACT** better preserves local structures and dimensions, while baselines can miss or distort fine details despite similar overall shapes.

Table 2: **Cumulative reconstruction ablation on oracle meshes.** Workflow stages are added sequentially to Codex (GPT-6); the final row is full **FACT**.

Condition	CD↓	SD-P95↓	F@1%↑	F@2%↑	Avg. SIoU↑	Min. SIoU↑
Codex	0.5402	1.0572	0.8783	0.9291	0.9888	0.9790
+Feature planning	0.1414	0.6836	0.9730	0.9950	0.9903	0.9813
+Measurement	0.0970	0.4878	0.9739	0.9926	**0.9950**	**0.9915**
+Refinement (**FACT**)	**0.0874**	**0.3740**	**0.9899**	**0.9990**	0.9902	0.9861

Component analysis. The cumulative ablation in Tab. 2 shows progressive reductions in CD and SD-P95 as feature planning, targeted measurement, and diagnostic refinement are added. Refinement further reduces SD-P95 by 23.3%, supporting the use of local feedback to correct residual surface discrepancies.

4.2 Contact Fidelity

Experimental setup. We evaluate collision proxies on articulated objects, using shared reference geometry, rigid-body assignments, and target tasks. All methods use CoACD (Wei et al., 2022). **Body** decomposes each rigid body as a whole, **Part** decomposes source parts within each body, and **Component** further separates closed connected components within each part. **DiagSearch** selects thresholds on the fixed Component partition using geometric, free-space, and motion diagnostics under hull-count budgets. **FACT** additionally revises local partitions and proxies using these diagnostics, followed by fidelity-constrained compression.

Evaluation protocol. Each object is evaluated on one lifting or articulation task with ten gripper trials. Interaction success rate (ISR) is the percentage of trials satisfying the task goal and pass contact, penetration, and non-target-motion checks. We also report volumetric IoU, task free-space occupancy (TFSO), and self-collision rate (SCR) over the articulated motion space. Total hull and vertex counts (H, V) measure proxy complexity, while mean physics-step time measures simulation cost.

Interaction fidelity and efficiency. **FACT** achieves 99.47% ISR versus 42.28% for DiagSearch, with 29.7% fewer hulls and 36.5% lower physics-step time (Tab. 3). It also improves IoU and TFSO, with no self-collisions at the articulated motion space.

Table 3: **Quantitative comparison of contact fidelity. FACT** achieves the lowest SCR and highest ISR with fewer hulls and lower physics-step time than DiagSearch.

Method	IoU (%)↑	SCR (%)↓	TFSO (%)↓	ISR (%)↑	Step time (ms)↓	H ↓	V ↓
Body	54.17	83.18	42.82	15.79	0.1035	76.1	9039.6
Part	68.75	79.52	21.35	28.42	0.1013	223.6	22123.7
Component	72.37	72.48	17.85	33.68	0.0956	306.2	21190.2
DiagSearch	74.50	59.78	16.15	42.28	0.0988	326.0	22043.4
FACT	**79.43**	**0.00**	**6.79**	**99.47**	**0.0627**	229.2	14493.7

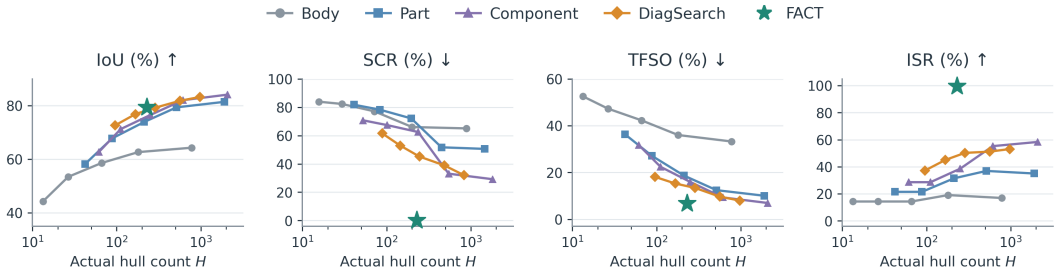

Figure 4: **Contact fidelity and proxy complexity.** Body, Part, and Component sweep CoACD thresholds; DiagSearch sweeps hull budgets. Failed settings are omitted; stars denote **FACT**.

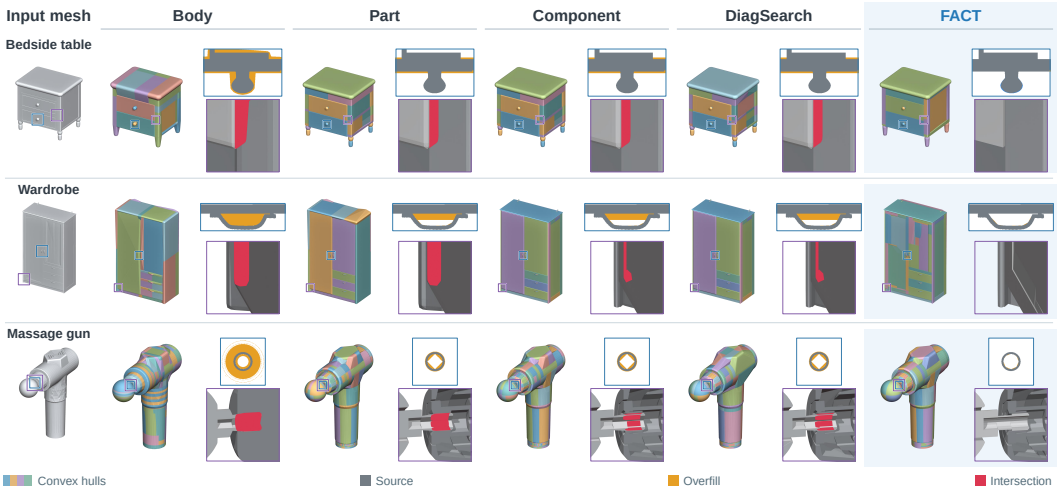

Figure 5: **Local collision proxy. FACT** preserves interaction-critical free space, including handle openings, while eliminating proxy-induced self-collisions.

Fidelity–complexity trade-off. The sweeps in Fig. 4 show that increasing hull count generally improves volumetric fit, but does not ensure reliable interaction. **FACT** achieves 99.47% ISR with 229.2 hulls on average, compared with 58.33% ISR and 2,081.5 hulls for the finest evaluated Component configuration. These results support targeted repair followed by fidelity-constrained compression rather than finer decomposition alone. Fig. 5 shows that **FACT** preserves the wardrobe's handle opening and removes proxy-induced intersections.

4.3 Dynamic Fidelity

Experimental setup. We evaluate dynamic prediction on held-out post-interaction clips of real objects. For each object, we first run the geometry and contact stages to obtain its reconstructed geometry G, kinematic model K, and refined collision proxies C. Reference trajectories are extracted from these videos in the recovered kinematic model's joint coordinates, and object dimensions are measured with a ruler. We compare **FACT** against five direct-inference baselines (Tab. 4): GPT-5.6 Sol (OpenAI, 2026a), GPT-6 Astra (OpenAI, 2026b), Gemini-3.8-Flash (Doshi & Popa, 2026), Qwen3.8-Max (Qwen Team, 2026), and Kimi-K3 (Team et al., 2026). These methods predict physical parameters from calibration videos or timestamped frames, whereas PhysX-Anything (Cao et al., 2026) uses the initial frame. All predictions are instantiated in MuJoCo (Todorov et al., 2012) with

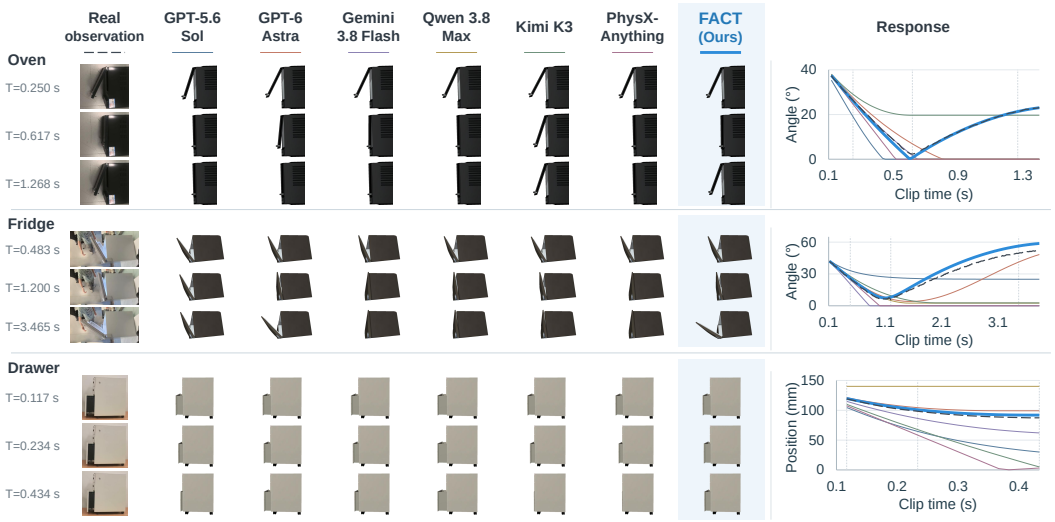

Figure 6: **Held-out motion prediction.** Time-aligned frames (left) and joint trajectories (right) compare real observations with simulated responses from **FACT** and the baselines for the oven, fridge, and drawer.

shared recovered geometry and joints, initial states, and solver settings. Response-model formulations follow each method's prediction procedure and may differ; the comparison therefore evaluates **FACT**'s full agentic dynamics calibration.

Evaluation protocol. Observed trajectories q_t and simulated trajectories $\hat{q}_t$ are compared on a shared physical-time grid. Let $R = q_{\max} - q_{\min}$ denote the fixed joint range and $A = \max_t q_t - \min_t q_t$ the observed motion range. We report velocity RMSE (VRMSE) normalized by R, trajectory MSE (TMSE) normalized by A^2, and the out-of-tolerance ratio OTR@10%, the percentage of samples with absolute position error exceeding $0.1R$. Velocities are computed from smoothed trajectories by numerical differentiation, while TMSE and OTR use the unsmoothed trajectories. Evaluation windows exclude initialization and near-stationary tails.

Held-out prediction. **FACT** achieves the lowest error across all three metrics (Tab. 4), reducing TMSE by 85.3% and OTR@10% by 59.2% relative to Gemini-3.8-Flash. Baseline performance varies across metrics, whereas our method consistently improves both velocity and positional agreement. Lower TMSE and OTR indicate that these gains extend beyond velocity matching to reproducing the trajectory on the shared physical-time grid. These results support the effectiveness of **FACT**'s agentic calibration loop. Fig. 6 shows that **FACT** captures the motion reversals of the oven and fridge doors, whereas several baselines miss these transitions.

Table 4: **Dynamic prediction on held-out clips.** **FACT** outperforms direct parameter inference across all three motion metrics.

Method	VRMSE↓	TMSE↓	OTR@10%↓
GPT-5.6 Sol	0.5181	0.6763	40.52
GPT-6 Astra	0.5853	1.1535	31.73
Gemini-3.8-Flash	0.5566	0.1711	27.05
Qwen3.8-Max	0.6101	0.4236	36.49
Kimi-K3	0.8541	0.6995	46.97
PhysX-Anything	0.7165	0.7549	32.01
FACT	**0.3786**	**0.0252**	**11.04**

5 Conclusion

We present **FACT**, an agentic framework for constructing articulated digital twins through quantitative measurements and simulation feedback. Its shared representation supports geometric reconstruction, task-aware collision refinement, and residual-guided dynamic calibration. Experiments show improved geometry, more reliable interaction with simpler proxies, and more accurate held-out motion prediction. **FACT** fixes geometry and kinematics during later refinement, limiting correction of upstream structural errors. Validation is restricted to selected interactions, sampled poses, and observed passive responses. Future work includes selective cross-stage revision and active acquisition of informative interactions.

AI USE DISCLOSURE

Generative AI tools were used to assist with writing and language polishing, and for literature retrieval and discovery. All AI-assisted content and references were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Hao Ai, Wenjie Chang, Jianbo Jiao, Ales Leonardis, and Eyal Ofek. Articulation in motion: Prior-free part mobility analysis for articulated objects by dynamic-static disentanglement. In *International Conference on Learning Representations*, volume 2026, pp. 87749–87774, 2026.

James Andrews. Navigation-driven approximate convex decomposition. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–9, 2024.

Mohamed el Amine Boudjoghra, Ivan Laptev, and Angela Dai. Unfoldart: Zero-shot recovery of full articulated 3d objects from text or image. *arXiv preprint arXiv:2606.30608*, 2026.

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-anything: Simulation-ready physical 3d assets from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5839–5848, 2026.

Peter Yichen Chen, Chao Liu, Pingchuan Ma, John Eastman, Daniela Rus, Dylan Randle, Yuri Ivanov, and Wojciech Matusik. Learning object properties using robot proprioception via differentiable robot-object interaction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5997–6004. IEEE, 2025.

Zoey Chen, Aaron Walsman, Marius Memmel, Kaichun Mo, Alex Fang, Karthikeya Vemuri, Alan Wu, Dieter Fox, and Abhishek Gupta. Urdformer: A pipeline for constructing articulated simulation environments from real-world images. *RSS*, 2024.

Bingquan Dai, Luo Li, Qihong Tang, Jie Wang, Xinyu Lian, Hao Xu, Minghan Qin, Xudong Xu, Bo Dai, Haoqian Wang, et al. Meshcoder: Llm-powered structured mesh code generation from point clouds. *Advances in Neural Information Processing Systems*, 38:8917–8954, 2026.

Tulsee Doshi and Raluca Ada Popa. Introducing gemini 3.8 flash and 3.8 flash cyber, September 2026. URL `https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/`.

Elisabetta Fedele, Boyang Sun, Leonidas Guibas, Marc Pollefeys, and Francis Engelmann. SuperDec: 3D Scene Decomposition with Superquadric Primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Egor Fokin and Manolis Savva. Visacd: Visibility-based gpu-accelerated approximate convex decomposition. In *47th Annual Conference of the European Association for Computer Graphics, Eurographics 2026 - Short Papers*, 2026.

Daoyi Gao, Yawar Siddiqui, Lei Li, and Angela Dai. Meshart: Generating articulated meshes with structure-guided transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7081–7091. IEEE, 2023.

Xincheng He, Rongrong Zhang, Wei Jiang, and Wenqiang Xu. Rigpi: Dynamic parameter identification of rigid body via vlm-seeded differentiable simulation. *arXiv preprint arXiv:2606.25212*, 2026.

Tao Hu, Jiaxin Ai, Licheng Wen, Xueheng Li, Shu Zou, Siqi Li, Nianchen Deng, Xinyu Cai, Hongbin Zhou, Pinlong Cai, et al. Itercad: An iterative multimodal agent for visually-grounded cad generation and editing. *arXiv preprint arXiv:2606.13368*, 2026.

Krishna Murthy Jatavallabhula, Miles Macklin, Florian Golemo, Vikram Voleti, Linda Petrini, Martin Weiss, Breandan Considine, Jerome Parent-Levesque, Kevin Xie, Kenny Erleben, Liam Paull, Florian Shkurti, Derek Nowrouzezahrai, and Sanja Fidler. gradsim: Differentiable simulation for system identification and visuomotor control. *International Conference on Learning Representations (ICLR)*, 2021.

Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. *ICCV*, 2025.

Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5606–5616. IEEE, 2022.

Zhao Jin, Zhengping Che, Tao Li, Zhen Zhao, Kun Wu, Yuheng Zhang, Yinuo Zhao, Zehui Liu, Qiang Zhang, Xiaozhu Ju, et al. Artvip: Articulated digital assets of visual realism, modular interaction, and physical fidelity for robot learning. In *International Conference on Learning Representations*, volume 2026, pp. 75272–75301, 2026.

R. Kenny Jones, Theresa Barton, Xianghao Xu, Kai Wang, Ellen Jiang, Paul Guerrero, Niloy J. Mitra, and Daniel Ritchie. Shapeassembly: Learning to generate programs for 3d shape structure synthesis. *ACM Transactions on Graphics (TOG)*, 39(6), 2020.

R Kenny Jones, David Charatan, Paul Guerrero, Niloy J Mitra, and Daniel Ritchie. Shapemod: Macro operation discovery for 3d shape programs. *ACM Transactions on Graphics (TOG)*, 40(4): 1–16, 2021.

R. Kenny Jones, Paul Guerrero, Niloy J. Mitra, and Daniel Ritchie. Shapecoder: Discovering abstractions for visual programs from unstructured primitives. *ACM Transactions on Graphics (TOG), Siggraph 2023*, 42(4), 2023.

R Kenny Jones, Paul Guerrero, Niloy Mitra, and Daniel Ritchie. Shapelib: Designing a library of programmatic 3d shape abstractions with large language models. *ACM Transactions on Graphics*, 2026.

Abhishek Joshi, Beining Han, Jack Nugent, Max Gonzalez Saez-Diez, Yiming Zuo, Jonathan Liu, Hongyu Wen, Stamatis Alexandropoulos, Karhan Kayan, Anna Calveri, Tao Sun, Gaowen Liu, Yi Shao, Alexander Raistrick, and Jia Deng. Procedural generation of articulated simulation-ready assets, 2025. URL https://arxiv.org/abs/2505.10755.

Soslan Kabisov, Vsevolod Kirichuk, Andrey Volkov, Marina Barannikov, Gennadiy Savrasov, Anton Konushin, Andrey Kuznetsov, and Dmitrii Zhemchuzhnikov. Cadreasoner: Iterative program editing for cad reverse engineering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6143–6153, June 2026.

Maksim Kolodiazhnyi, Denis Tarasov, Dmitrii Zhemchuzhnikov, Alexander Nikulin, Ilya Zisman, Anna Vorontsova, Anton Konushin, Vladislav Kurenkov, and Danila Rukhovich. cadrille: Multimodal cad reconstruction with reinforcement learning. *arXiv preprint arXiv:2505.22914*, 2025.

Jens U Kreber and Joerg Stueckler. Guiding diffusion-based articulated object generation by partial point cloud alignment and physical plausibility constraints. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3206–3214. IEEE, 2025.

Long Le, Jason Xie, William Liang, Hung-Ju Wang, Yue Yang, Yecheng Jason Ma, Kyle Vedder, Arjun Krishna, Dinesh Jayaraman, and Eric Eaton. Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. In *International Conference on Learning Representations*, volume 2025, pp. 17578–17602, 2025.

Jiahui Lei, Congyue Deng, Bokui Shen, Leonidas Guibas, and Kostas Daniilidis. Nap: Neural 3d articulated object prior. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31878–31894, 2023.

Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. PAC-neRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhe Li, Xiang Bai, Jieyu Zhang, Zhuangzhe Wu, Che Xu, Ying Li, Chengkai Hou, and Shanghang Zhang. Urdf-anything: Constructing articulated objects with 3d multimodal language model. *Advances in Neural Information Processing Systems*, 38:94974–95002, 2026.

Youtian Lin, Yikang Yang, Zhanpeng Hu, Mengqi Zhou, Feihu Zhang, Xun Cao, Jiaheng Liu, and Yao Yao. Procedura: Agentic 3d modeling with procedural control. *arXiv preprint arXiv:2608.26238*, 2026.

Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 352–363, 2023.

Jiayi Liu, Hou In Ivan Tam, Ali Mahdavi-Amiri, and Manolis Savva. Cage: Controllable articulation generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17880–17889, June 2024.

Jiayi Liu, Denys Iliash, Angel X Chang, Manolis Savva, and Ali Mahdavi-Amiri. SINGAPO: Single image controlled generation of articulated parts in object. *ICLR*, 2025a.

Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Building interactable replicas of complex articulated objects via gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Haozhe Lou, Mingtong Zhang, Haoran Geng, Hanyang Zhou, Sicheng He, Zhiyuan Gao, Siheng Zhao, Jiageng Mao, Pieter Abbeel, Jitendra Malik, et al. D-rex: Differentiable real-to-sim-to-real engine for learning dexterous grasping. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuzhe Luo, Zherong Pan, Kui Wu, Xingyi Du, Yun Zeng, Xiangjun Tang, Yiqian Wu, Xiaogang Jin, and Xifeng Gao. Rl-acd: Reinforcement learning-based approximate convex decomposition. *ACM Trans. Graph.*, 44(6), 2025. ISSN 0730-0301.

Khaled Mamou, E Lengyel, and A Peters. Volumetric hierarchical approximate convex decomposition. *Game engine gems*, 3:141–158, 2016.

OpenAI. Introducing codex. `https://openai.com/index/introducing-codex/`, 2025.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026a. URL `https://openai.com/index/gpt-5-6/`.

OpenAI. GPT-6 Astra: A New Generation of Intelligence, 2026b. URL `https://openai.com/index/gpt-6-astra/`.

Nicholas Pfaff, Evelyn Fu, Jeremy Binagia, Phillip Isola, and Russ Tedrake. Scalable real2sim: Physics-aware asset generation via robotic pick-and-place setups. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6296–6303. IEEE, 2025.

Xiaowen Qiu, Jincheng Yang, Yian Wang, Zhehuan Chen, Yufei Wang, Tsun-Hsuan Wang, Zhou Xian, and Chuang Gan. Articulate anymesh: Open-vocabulary 3d articulated objects modeling. *arXiv preprint arXiv:2502.02590*, 2025.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Danila Rukhovich, Elona Dupont, Dimitrios Mallis, Kseniya Cherenkova, Anis Kacem, and Djamila Aouada. Cad-recode: Reverse engineering cad code from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9801–9811, 2025.

Jiayi Su, Youhe Feng, Zheng Li, Jinhua Song, Yangfan He, Botao Ren, and Botian Xu. Artformer: Controllable generation of diverse 3d articulated objects. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1894–1904. IEEE, 2025.

Gabriel Tavernini, Elisabetta Fedele, Tiago Novello, Leonidas Guibas, Marc Pollefeys, and Francis Engelmann. Superflex: Deformable superquadrics for point cloud decomposition. In *European Conference on Computer Vision*, pp. 457–474. Springer, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Daniel Thul, L'ubor Ladický, Sohyeon Jeong, and Marc Pollefeys. Approximate convex decomposition and transfer for animated meshes. *ACM Trans. Graph.*, 37(6), 2018. ISSN 0730-0301.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Penghao Wang, Siyuan Xie, Hongyu Yan, Xianghui Yang, Jingwei Huang, Chunchao Guo, and Jiayuan Gu. Artllm: Generating articulated assets via 3d llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34281–34291, 2026a.

Xinjie Wang, Liu Liu, Taojun Ding, Andrew Choi, Chaodong Huang, Mengao Zhao, Ziang Li, Jackson Jiang, Chunlei Yu, Shengxiang Liu, et al. Embodiedgen v2: An agentic, simulation-ready 3d world engine for embodied ai. *arXiv preprint arXiv:2607.07459*, 2026b.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022.

Yijia Weng, Bowen Wen, Jonathan Tremblay, Valts Blukis, Dieter Fox, Leonidas Guibas, and Stan Birchfield. Neural implicit representation for building digital twins of unknown articulated objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3141–3150. IEEE, 2024.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yuezhi Yang, Qixing Huang, Mikaela Angelina Uy, and Nicholas Sharp. Learning convex decomposition via feature fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Yunhan Yang, Chunshi Wang, Junliang Ye, YANG LI, Zanxin Chen, Zehuan Huang, Yao Mu, Zhuo Chen, Chunchao Guo, and Xihui Liu. Physforge: Generating physics-grounded 3d assets for interactive virtual world. In *Forty-third International Conference on Machine Learning*, 2026b.

Jingwen Ye, Yuze He, Yanning Zhou, Yiqin Zhu, Kaiwen Xiao, Yong-Jin Liu, Wei Yang, and Xiao Han. Primitiveanything: Human-crafted 3d primitive assembly generation with auto-regressive transformer. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions On Graphics (TOG)*, 43(4):1–20, 2024.

Longwen Zhang, Qixuan Zhang, Haoran Jiang, Yinuo Bai, Wei Yang, Lan Xu, and Jingyi Yu. Bang: Dividing 3d assets via generative exploded dynamics. *ACM Transactions on Graphics (TOG)*, 44 (4):1–21, 2025.

Shuyuan Zhang, Chenhan Jiang, Zuoou Li, and Jiankang Deng. Shapecraft: Llm agents for structured, textured and interactive 3d modeling. *Advances in Neural Information Processing Systems*, 38:65116–65144, 2026.

Mandi Zhao, Yijia Weng, Dominik Bauer, and Shuran Song. Real2code: Reconstruct articulated objects via code generation. In *International Conference on Learning Representations*, volume 2025, pp. 668–686, 2025.

Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. *European Conference on Computer Vision (ECCV)*, 2024.

Matt Zhou, Ruining Li, Xiaoyang Lyu, Zhaomou Song, Zhening Huang, Chuanxia Zheng, Christian Rupprecht, Andrea Vedaldi, and Shangzhe Wu. Articraft: An agentic system for scalable articulated 3d asset generation. *arXiv preprint arXiv:2605.15187*, 2026.

A Data and Experimental Settings

Geometry evaluation (Tab. 1–2) uses all 21 reference assets. Contact evaluation (Tab. 3) uses a 19-object subset; `grand_piano` and `oven_6_333` are excluded because extensive non-manifold geometry prevents reliable construction of the closed-material references required by the contact metrics. The dynamics evaluation (Tab. 4) uses six separate real objects. Each real object is first reconstructed through the geometry and contact stages, and the resulting articulated twin is then used for dynamics calibration and evaluation.

A.1 Reference assets

The reference collection contains 13 ArtVIP (Jin et al., 2026) assets and eight artist-created commercial assets. Tab. 5 lists the assets and their reviewed articulation structures. Geometry evaluation uses all 21 assets and the complete reference surfaces, including interior geometry. Contact evaluation uses 19 assets and reviewed closed material units, excluding geometry marked as visual-only. Objects with a single kinematic body are excluded from SCR, for which self-collision is not applicable. Repairs to open boundaries and material solids, as well as reviewed body assignments and joints, are shared across all contact methods.

Table 5: Reference dataset for the geometry and contact stages. V: ArtVIP; A: commercial artist asset.

Asset ID	Source	Kinematic bodies
bakingchamber	V	2
bedsidetable	V	3
brimnes_cabinet	V	3
cabinet_1	V	3
drawer_kit_3	V	6
grand_piano	A	8
handheld_vacuum_cleaner	A	1
humidifier	A	1
lipstick	A	2
massage_gun	A	2
match	V	2
microwave_1	V	3
motor	A	2
musken_wardrobe	V	6
oven_6_333	V	6
pet_feeder	A	1
power_bank	A	1
pressure_pump_2	V	2
refrigerator_8	V	7
rolling_washer_5	V	3
washing_machine_2	V	4

A.2 Real-object recordings

Tab. 6 summarizes the six real objects and their recording coverage. The reported clip count is the total number of recorded clips for each object. For each object, one or two clips are held out exclusively for evaluation, while all remaining clips are available for observation extraction, response-model selection, parameter fitting, and stopping decisions. Recordings are 1280×720 at approximately 60 FPS. A measured exterior dimension determines the scale factor $s = L_{\text{real}}/L_{\text{mesh}}$. The resulting uniform scale is applied consistently to visual and collision geometry, joint locations, and prismatic travel, while revolute joint angles remain unchanged.

Table 6: Real-object inventory and recording coverage.

Object	Observed motion	Kinematic bodies	Duration (s)	Clips
Chest	Lid with passive stays	4	25.84	2
Drawer	Lower drawer	4	21.93	6
Trashcan	Coupled lid and pedal	3	18.91	2
Cabinet	Middle door	4	17.08	3
Fridge	Upper door	4	41.70	7
Oven	Vertical-axis door	2	32.47	9

B Geometry Fidelity

B.1 Implementation and validation

Feature implementation. Feature plans associate editable features with source regions while preserving node identities. After applying scene-node transforms, GLB/glTF coordinates are mapped to Blender as $(x, y, z) \mapsto (x, -z, y)$; OBJ/STL coordinates are unchanged. Coordinate frames are defined from observed geometric and semantic datums. Parameter records track measurements, derived constraints, agent adjustments, defaults, and user inputs together with their source regions, frames, and measurement settings. Contour reduction and section sampling are selected per feature type. Reconstructed meshes are generated from feature parameters rather than copied source triangles, and exported part and feature identifiers preserve assembly correspondences.

Validation. Let D_M denote the source bounding-box diagonal. Automated validation requires per-axis dimension errors below 5%, center error below $0.02D_M$, both directed global P95 surface distances below $0.02D_M$, and canonical-view silhouette IoU of at least 0.90. All checks are performed in source assembly coordinates without alignment. When source-part correspondence is available, part-level checks use 2,048 samples per direction and a threshold of $0.02D_k$, where D_k is the source-part diagonal. For unsegmented inputs, recovered regions are used only for visual diagnosis and are not treated as source-part ground truth. Numerical validation is supplemented by structural inspection, with intentional simplifications recorded as residuals.

Table 7: Inference configurations for learned geometry baselines.

Method	Configuration
CAD-Recode	256 input points; 10 generated programs; CD-based selection among executable candidates
MeshCoder	16,384 normalized input points; generated Blender code executed without manual repair
PrimitiveAnything	10,000 surface points with normals sampled from the normalized input mesh; official autoregressive inference pipeline
SuperFlex	4,096-point input point cloud; feed-forward deformable-superquadric checkpoint; no object-specific optimization

B.2 Inputs and baseline configurations

For the image-input setting, Codex and Articraft receive only the input image, while FACT incorporates Rodin generation into its pipeline, producing an original or part-split asset that is subsequently frozen for reconstruction. For oracle-mesh evaluation, learned baselines receive the whole-object geometry without part grouping, while agentic methods retain the scene hierarchy and may inspect individual parts. All agentic geometry methods use GPT-6 Astra with High reasoning level. Codex operates through Blender MCP without the FACT reconstruction workflow, while Articraft follows its native pipeline. Inference configurations for the learned baselines are summarized in Tab. 7.

B.3 Geometry evaluation metrics

Pose and alignment. All metrics are computed on whole-object visual surfaces in a canonical pose. Articraft URDFs apply the visual-mesh scales and link/joint transforms, while FACT assets use their closed or default pose. Let c_P, c_S denote the AABB centers of the candidate and reference, and let ℓ_P, ℓ_S denote their longest AABB side lengths. The candidate is aligned by the uniform transform

$$p' = c_S + \frac{\ell_S}{\ell_P}(p - c_P).$$

No metric-driven rotation search, ICP, nonuniform scaling, or per-part alignment is applied. The resulting scores therefore measure normalized shape and relative assembly dimensions, rather than absolute scale or global translation.

Surface distances and F-scores. Let S denote the reference surface, P the aligned candidate surface, and D_S the diagonal of the reference AABB. We draw $n = 4{,}096$ area-weighted surface samples in each direction, excluding nonfinite triangles and triangles with area at most 10^{-15}. Distances are unsquared Euclidean point-to-surface distances. For $x_i \sim S$ and $y_j \sim P$, define

$$a_i = \frac{\min_{p \in P} \|x_i - p\|_2}{D_S}, \qquad b_j = \frac{\min_{s \in S} \|y_j - s\|_2}{D_S}.$$

We report

$$\mathrm{CD} = \frac{100}{2n}\left(\sum_i a_i + \sum_j b_j\right), \qquad \text{SD-P95} = 100\, Q_{0.95}(b_1, \ldots, b_n),$$

where $Q_{0.95}$ uses linear percentile interpolation. SD-P95 measures reconstruction-to-reference error only; missing reference geometry instead contributes to the reverse CD term and F-score recall. For $\tau \in \{0.01, 0.02\}$,

$$p_\tau = \frac{1}{n}\sum_j \mathbf{1}[b_j \le \tau], \qquad r_\tau = \frac{1}{n}\sum_i \mathbf{1}[a_i \le \tau], \qquad F_\tau = \frac{2p_\tau r_\tau}{p_\tau + r_\tau},$$

with $F_\tau = 0$ when $p_\tau + r_\tau = 0$. F@1% and F@2% are reported on $[0, 1]$.

Silhouette IoU. We rasterize triangle projections into binary 512×512 masks using Pillow without antialiasing. The three orthographic projections are $(-x, z)$, (y, z), and (x, y). Each candidate–reference pair shares the union of their projected bounds, a common isotropic pixel scale, and a 31-pixel margin. Avg. SIoU and Min. SIoU denote the mean and minimum IoU across the three views, respectively.

B.4 Case Study: Baking-chamber Handle Refinement

The baking chamber contains eight parts represented by 15 generated meshes. In source coordinates, the agent extracts a handle section from the 19 unique vertices satisfying $|z + 0.240055| < 10^{-5}$. Section extrema determine the center c_x, half-width h_x, front coordinate y_f, and depth d. The agent selects a half-superellipse parameterization,

$$u_i = \frac{|x_i - c_x|}{h_x}, \qquad v_i = \frac{y_i - y_f}{d},$$

and the numerical tool fits the superellipse exponent as

$$e^\star = \arg\min_{1.2\le e\le 5} \sum_i \left(u_i^e + v_i^e - 1\right)^2.$$

The resulting parameters are $c_x \approx 0.02538$, $h_x \approx 0.06353$, $y_f \approx -0.48697$, $d \approx 0.06631$, and $e^\star \approx 2.6593$. The agent then revises the feature representation to loft height-varying half-superellipse sections while preserving the observed open mating face.

Although the initial reconstruction passed whole-object validation, part-level diagnostics revealed an inaccurate handle profile and spurious closure surfaces (Tab. 8). The agent therefore remeasures the handle and revises the affected feature representations. After revision, all eight parts satisfy the local threshold, despite a small increase in the global reconstruction-to-source P95 distance.

Table 8: Local baking-chamber refinement. Surface distances are the maximum of the two directed part P95 values, divided by D_k.

Part	Before	After	Revision
Handle trim	0.037068	0.009469	Remove rear closure
Handle	0.087434	0.009444	Remeasure; replace extrusion with open loft
Dark surround	0.114964	0.002874	Remove concave mating surface

C Contact Fidelity

C.1 Proxy Initialization, Repair, and Compression

Proxy Initialization. Within each rigid body, the reference material is the union of reviewed solids, with overlapping volume counted once. Repartitioning preserves source-face ownership and never crosses body boundaries. The initial proxy C^0 is the Component decomposition: closed, consistently oriented face-connected components are separated within each source part, while residual components remain grouped. The object-level hull budget is fixed as $H_{\max} = H(C^0)$.

CoACD 1.0.14 uses threshold 0.05, automatic preprocessing, preprocessing resolution 50, resolution 2000, and MCTS nodes/iterations/depth of 20/150/3. We retain the default merging behavior, disable PCA, decimation, and extrusion, use the 256-vertex setting and convex-hull approximation, and impose no per-call hull cap. Body, Part, and Component use the default threshold; DiagSearch uses its frozen selection under a 512-hull budget, whereas FACT uses the object-specific budget $H_{\max}$.

Repair search. The agent proposes local repartitions and regional decomposition settings, while the numerical backend regenerates and evaluates the affected proxies. A candidate is feasible only if it satisfies the hull budget $H(C) \leq H_{\max}$, passes the sampled motion checks, and keeps the prescribed volumetric and free-space errors within their allowed increases relative to C^0. Candidates violating any of these constraints are rejected. Among feasible candidates, the search first minimizes worst-region contact P95; candidates within $0.0005D$ of the best value are compared by development loss, and those within 0.005 of the best development loss are finally ranked by proxy complexity (H, V), where D is the contact-reference diagonal. The development loss equally averages $1 - \mathrm{IoU}$, ROI occupancy, and motion error, with the motion term omitted for static objects. The selected feasible proxy is fixed as the repair anchor C^a.

Compression. All compression candidates are evaluated against the fixed anchor C^a, rather than the preceding iterate. For each body, the allowed increases in $(1-\mathrm{IoU}, m, x)$ are $(0.02, 0.002, 0.05)$, and each eligible development ROI permits an occupancy increase of at most 0.01. Auxiliary source-normal guards allow P95 and maximum displacement to increase by at most $0.0005D$ and $0.002D$, respectively, with no increase in missing normal intersections.

C.2 Contact Diagnostics and Metrics

Let D denote the reference object's bounding-box diagonal.

Volumetric fidelity. For reference material G_b and proxy union C_b of body b, we measure

$$\mathrm{IoU}_b = \frac{|G_b \cap C_b|}{|G_b \cup C_b|}, \qquad m_b = \frac{|G_b \setminus C_b|}{|G_b|}, \qquad x_b = \frac{|C_b \setminus G_b|}{|G_b|},$$

where m_b and x_b are the missing- and excess-volume ratios, respectively. Per-object values are averaged equally across bodies. Volume estimation uses eight scrambled Sobol batches with 2^{15} development or 2^{17} evaluation queries. Samples combine uniform sampling in a source-derived domain, initially padded by $0.1D$, with volume-proportional sampling from reference-material PCA boxes; inverse-density weights correct for overlapping boxes. All methods share the same samples and reference labels, which are recomputed if the sampling domain expands. Reference occupancy is determined by three signed-ray tests with tolerance $10^{-9}D$. Ambiguous queries define lower and upper IoU bounds, and we report their midpoint. Missing and excess ratios use only queries with known reference labels; uncertainty from ambiguous queries and across Sobol batches is tracked separately.

Self-collision. For each sampled motion path, let Q^{free} contain poses at which the reference bodies are classified as collision-free. We compute

$$\mathrm{SCR} = \frac{100}{|Q^{\mathrm{free}}|} \sum_{q \in Q^{\mathrm{free}}} I_\epsilon(C, q),$$

where $I_\epsilon(C, q) = 1$ if an interbody hull intersection contains a ball of radius $\epsilon = 10^{-4}D$, tested using inward-offset halfspaces. Same-body overlaps are ignored. Evaluation poses are densely sampled throughout the prescribed articulated motion space of each object. Rates are averaged equally over eligible motion paths, available decomposition seeds, and the 15 articulated contact objects.

Task free-space occupancy. For region r at pose q with source-defined probe distribution $\mu_{r,q}$, we measure

$$\mathrm{TFSO}_{r,q} = 100 \Pr_{p \sim \mu_{r,q}} \left[p \in \bigcup_b T_b(q) C_b \,\middle|\, p \text{ is confirmed reference-free} \right],$$

where $T_b(q)$ denotes the transform of body b. Free probes are determined from the full reference-material union independently of the candidate proxy. Access probes are sampled from source surfaces by area with offsets in $[10^{-4}D, 0.02D]$, while motion-clearance probes use $0.01D$ bands near neighboring bodies. A region is eligible only with at least 128 confirmed-free samples overall and 16 per batch. Scores are averaged equally across batches, eligible poses, regions within each access or clearance category, and available categories.

C.3 INTERACTION PROTOCOL

Each object is assigned one prescribed interaction task, executed by an actuated Cartesian gripper with wrist rotation and opposing fingers or by a pressing pad. Object joints remain passive and move only through physical contact. Task trajectories use cubic smoothstep interpolation, $3u^2 - 2u^3$, with normalized time $u \in [0, 1]$. Controller parameters and task durations are fixed across methods. Each proxy is evaluated over ten trials: trial 00 uses the nominal setup, while trials 01–09 apply uniformly sampled position offsets with half-ranges $(1, 0.4, 0.4)$ mm. The same offsets are used across methods, with no orientation perturbation.

Success criteria. Articulation tasks must maintain at least 90% of the target progress throughout the final hold. Static lifting tasks require a center-of-mass rise of at least 90 mm and a source-to-floor clearance of at least 20 mm throughout the sampled hold states. All trials additionally require penetration of at most 1 mm and tracking error of at most 20 mm; where applicable, non-target joint displacement must remain below 0.03 m for prismatic joints or 0.03 rad for revolute joints. Forbidden non-target contacts and solver warnings constitute failure. All trials run for the prescribed horizon, and failure to reach or maintain the task goal is counted as unsuccessful.

C.4 SIMULATION AND TIMING

Contact evaluation uses MuJoCo 3.13.0 with a 1 ms timestep, `implicitfast` integration, a Newton solver with 80 iterations, an elliptic friction cone, and gravity $(0, 0, -9.81)$ m/s^2. The solver tolerance is 10^{-9}, except 10^{-10} for the cabinet. Contact parameters are `solref=(0.002,1)`, `solimp=(0.95,0.99,0.001)`, `condim=4`, zero margin, and friction $(0.8, 0.005, 0.0001)$. Within each object, body inertials, joint properties, contact masks, and controllers are fixed across methods.

Physics-step timing is measured on an Intel Core i9-14900K using one serial pass per proxy. Native `mj_step` calls are timed after a 0.5 s warm-up; control, rendering, model loading, and source-geometry audits are excluded. Reported step times are averaged over simulation steps, decomposition seeds, and objects.

C.5 MICROWAVE REFINEMENT TRACE

The initial shell proxy induces false collisions with the door and turntable. The agent repartitions the shell along measured cavity planes into six structural regions. It then sets the CoACD thresholds for the door frame and shell to 0.01 and 0.025, respectively, with preprocessing disabled. A 64-vertex turntable candidate passes the sampled motion checks but is rejected for 45.23% missing material; restoring the thin-disc boundary with 100 vertices resolves this failure. Tab. 9 reports the corresponding development and final measurements.

Starting from the repaired anchor C^a, compression removes 82 hulls. The largest per-body increases in $(1 - \mathrm{IoU}, m, x)$ are approximately $(0.003254, 0.001529, 0.001767)$, and the largest ROI occupancy increase is 0.005859, all within the prescribed fidelity bounds. Subsequent local edits either violate these guards or fail to further reduce the hull count, yielding a local stopping point. The final proxy succeeds in all ten interaction trials while retaining bounded approximation error.

Table 9: Microwave refinement trace from the initial proxy C^0 to the repaired anchor C^a and compressed proxy C^{final}. IoU and ROI occupancy are percentages; fidelity constraints are enforced per body and region rather than on aggregate values. False/free reports proxy-induced collision poses among reference-free samples.

Stage / samples	$H \downarrow$	$V \downarrow$	IoU $\uparrow$	ROI occ. $\downarrow$	False/free $\downarrow$
C^0 / development	420	19,641	95.6736	24.4141	93/93
C^a / development	364	14,449	99.7027	16.9789	0/93
C^{final} / development	282	12,385	99.5943	17.0410	0/93
C^{final} / final	282	12,385	99.5784	16.9522	0/603

D DYNAMIC FIDELITY

We evaluate passive-response prediction on six reconstructed real objects. For each object, one or two clips are reserved exclusively for evaluation, while all remaining clips are used for calibration. The same split is used for FACT and all baselines: model construction or parameter inference uses only the calibration clips, and the held-out clips are used only for final evaluation (Section A.2).

D.1 TRAJECTORY EXTRACTION

Timestamped videos are converted to joint-coordinate trajectories in the reconstructed kinematic model. We first identify hand-free intervals and define release as the first reliable passive frame. Object motion is recovered from image-space features, including tracked points, silhouettes, and edges, and mapped to the corresponding joint coordinates using geometry-aware planar mappings or pivot constraints. Measurements are filtered by detector confidence and local temporal consistency; isolated gaps are interpolated only when supported by reliable neighboring observations, while unresolved detections are masked. Evaluation uses the accepted, unsmoothed position measurements. All methods share the same trajectory references, coordinate mappings, masks, and evaluation windows.

D.2 RESPONSE MODELS

Let q denote the joint coordinate defined by the reconstructed kinematic model, oriented positively toward opening, and let $v = \dot{q}$. From the reconstructed mechanism and observed passive motion, the agent selects a response-model family, diagnoses systematic residuals, and revises the model when needed. Numerical tools then fit the free physical parameters of the selected model. Geometry, kinematics, collision proxies, inertial estimates, and fixed boundary conditions remain unchanged during each fit.

Tab. 10 summarizes the final agent-constructed response models. These models capture the dominant passive effects observed across the six mechanisms, including dry and viscous resistance, state-dependent restoring terms, piecewise spring–damper behavior, and impact-like closure responses.

Table 10: Agent-constructed response models for dynamics calibration. Reported values are the final fitted parameters.

Object	Response model and fitted parameters
Chest	Gravity-driven linkage with Coulomb lid resistance, $\tau_{\text{res}} = -\tau_c \operatorname{Sign}(v)$. *Fit:* $\tau_c = 0.56757\ \text{N m}$.
Drawer	Viscous and Coulomb resistance, $\ddot{q} = -\beta v - \alpha \operatorname{Sign}(v)$. *Fit:* $\beta = 4.04196\ \text{s}^{-1}$, $\alpha = 0.375197\ \text{m/s}^2$.
Trashcan	Coupled lid–pedal mechanism with viscous lid resistance, $\tau_{\text{res}} = -bv$. *Fit:* $b = 0.00790463\ \text{N m s/rad}$.
Cabinet	Piecewise spring–damper response with distinct closing, intermediate, and opening regimes. *Fit:* $(k_c, b_c) = (0.19321, 0.34609)$, $(k_o, b_o) = (11.4396, 0.505714)$, $q_o = 89.8416^\circ$.
Fridge	State-dependent restoring response with smooth velocity resistance, $\ddot{q} = Ae^{-q/w} - f \tanh(v/0.025)$. *Fit:* $A = 21.8265$, $w = 0.064639$, $f = 0.212000$.
Oven	Coulomb deceleration with state-triggered closure restitution, $\ddot{q} = -a \operatorname{Sign}(v)$. *Fit:* $a = 0.842825\ \text{rad/s}^2$, $e = 0.734856$.

For the cabinet, the final piecewise response is

$$\tau(q, v) = \begin{cases} -k_c q - b_c v, & q < q_1, \\ -b_m v, & q_1 \le q \le q_2, \\ k_o(q_o - q) - b_o v, & q > q_2, \end{cases}$$

where the transition locations and intermediate damping are fixed during parameter fitting. For the fridge, the fitted exponential restoring term is combined with a fixed state-dependent gate during replay. For the oven, reaching the closure boundary triggers a restitution update determined by the fitted coefficient e. These fixed transition and boundary rules define the response family selected by the agent but are not themselves optimized in the final fit.

D.3 Parameter fitting

We instantiate the objective in Eq. (2) as a normalized weighted least-squares problem. For accepted observations $\mathcal{T}_k$, shared physical parameters θ, and sequence-specific initial states $\xi_k = (q_{0,k}, v_{0,k})$, we solve

$$\min_{(\theta,\{\xi_k\})\in\Omega} \frac{1}{2} \sum_k \left[c_k \sum_{i\in\mathcal{T}_k} \left(\frac{q_k^{\mathrm{sim}}(t_i;\theta,\xi_k) - q_{ki}^{\mathrm{obs}}}{s_q} \right)^2 + R_k(\xi_k) \right].$$

Here, s_q normalizes trajectory residuals, c_k controls sequence weighting, and R_k optionally regularizes uncertain initial states. Geometry, kinematics, inertial estimates, camera mappings, and time offsets remain fixed during fitting.

Table 11: Candidate parameter and initial-state bounds considered during calibration and model revision. Final selected models may retain only a subset of the listed parameters.

Object	Parameter bounds	Initial-state bounds
Chest	$\tau_c \in [0.05, 1.5]$, $b \in [0, 0.12]$.	$q_0 : \pm 1.3^\circ$; $v_0 : \pm r_v$, with $v_0 \le 0$.
Drawer	$\beta \in [0.01, 30]$, $\alpha \in [0, 10]$.	$q_0 : \pm 0.004\,\mathrm{m}$; $v_0 : \pm \max(0.3\,\mathrm{m/s}, 0.35\lvert\widetilde{v}_0\rvert)$.
Trashcan	$b \in [0, 0.04]$, $\tau_c \in [0, 0.035]$.	$q_0 : \pm 3^\circ$; $v_0 : \pm 57.3^\circ/\mathrm{s}$, with $v_0 \le 0$.
Cabinet	$k_c \in [0.001, 2]$, $b_c \in [0.001, 3]$, $k_o \in [0.1, 40]$, $b_o \in [0.005, 3]$, $q_o \in [86^\circ, 97^\circ]$.	$q_0 : \pm 2^\circ$; $v_0 : \pm 20^\circ/\mathrm{s}$.
Fridge	$A \in [0.1, 200]$, $w \in [0.015, 0.25]$, $f \in [0.01, 1.5]$.	$q_0 : \pm 1.5^\circ$; $v_0 : \pm 8^\circ/\mathrm{s}$.
Oven	$a \in [0.03, 8]$, $b \in [0, 8]$, $e \in [0.05, 0.95]$, $\delta \in [0, 4^\circ]$.	$q_0 : \pm 2^\circ$, with $q_0 \ge 0$; $v_0 : \pm \max(20^\circ/\mathrm{s}, 0.25\lvert\widetilde{v}_0\rvert)$.

Initial states are estimated from short release windows using low-order polynomial fits and, when uncertain, optimized within bounded neighborhoods of these estimates. The numerical backend uses bounded nonlinear least squares with model-specific forward evaluators: direct simulation when linkage dynamics are retained, analytical responses when available, and numerical integration otherwise. Multiple initializations are evaluated when needed, and the lowest-cost feasible solution is retained. Tab. 11 reports the parameter bounds considered during calibration and model revision, together with the initial-state bounds. The final selected response model may retain only a subset of these candidate parameters.

Residual-guided model revision. After each fit, the agent inspects trajectory residuals to determine whether errors arise from parameter values, uncertain initial conditions, or an inadequate response-model form. It then revises the optimized variables, bounds, or model structure and refits numerically. When multiple response models explain the observations, we retain the simplest model satisfying the prescribed per-sequence error criteria.

D.4 Replay and evaluation

Baseline inputs. All direct-inference baselines operate only on the calibration observations. Language-model baselines receive calibration motion videos or timestamped frames in the modality supported by their respective interfaces, while PhysX-Anything receives an initial frame from the calibration recordings. Their inferred response models or physical parameters are then fixed and replayed on the held-out evaluation clips. No baseline receives the evaluation trajectories during model or parameter inference. Because the baselines differ in input modality and response-model param-

eterization, the comparison evaluates end-to-end motion prediction rather than enforcing identical intermediate representations.

Replay protocol. All methods are evaluated using the same reconstructed geometry, joint coordinates, observation-derived initial states, and numerical settings within each object. Initial states are estimated from a short release window and are not fitted to the evaluation trajectory. Method-specific response models and inferred physical parameters are otherwise preserved. Forward replay uses MuJoCo 3.13.0 with Euler integration, Newton solving, a dense Jacobian, elliptic friction cones, tolerance 10^{-10}, gravity $(0, 0, -9.81)$ m/s^2, and zero external input after release.

Temporal alignment and scoring windows. Observed and simulated trajectories are linearly interpolated onto a common physical-time grid at approximately 60 Hz. Position metrics use the accepted unsmoothed observations, while velocities are obtained using a shared smoothing and numerical-differentiation procedure. Missing or unreliable observations are masked consistently. Evaluation windows and stationary-tail truncation are determined from the reference trajectory alone and then shared across methods.

Motion metrics. Let $\mathcal{V}_q$ and $\mathcal{V}_v$ denote the valid position and velocity indices, $R = q_{\max} - q_{\min}$ the fixed joint range, and

$$A = \max_{i \in \mathcal{V}_q} q_i - \min_{i \in \mathcal{V}_q} q_i$$

the observed motion amplitude. We report

$$\mathrm{VRMSE} = \frac{1}{R}\sqrt{\frac{1}{|\mathcal{V}_v|}\sum_{i \in \mathcal{V}_v}(\widehat{v}_i - v_i)^2},$$

$$\mathrm{TMSE} = \frac{\sum_{i \in \mathcal{V}_q}(\widehat{q}_i - q_i)^2}{|\mathcal{V}_q| A^2},$$

$$\mathrm{OTR@10\%} = \frac{100}{|\mathcal{V}_q|}\sum_{i \in \mathcal{V}_q} \mathbf{1}\left[|\widehat{q}_i - q_i| > 0.1R\right].$$

Metrics are averaged equally across coordinates, clips, and objects, in that order.